

# Fine-grained Long-range Workload Prediction

Ivan, Petru, and Zebo

October 11, 2016

Inception

# Embedded Learning™

- Learning is key
- Learning requires data
- Real data are expensive
- Simulated data are expensive

# Synthesize to Excel

1. Synthesize data
2. Explore wildly your wild ideas
3. Prototype the promising ones
4. Fine-tune in a real environment

# Synthesize to Excel

- "How accurate...?"
- "How to validate...?"
- "How to guarantee...?"

**Real Data**

# Google's Cluster Data

- 1 month, May 2011
- 13,000 machines
- 700,000 jobs
- 1,000 users

# Google's Cluster Data

- Machine events
- Machine attributes
- **Job events**
- Task events
- Task constraints
- **Task resource usage**



# Question

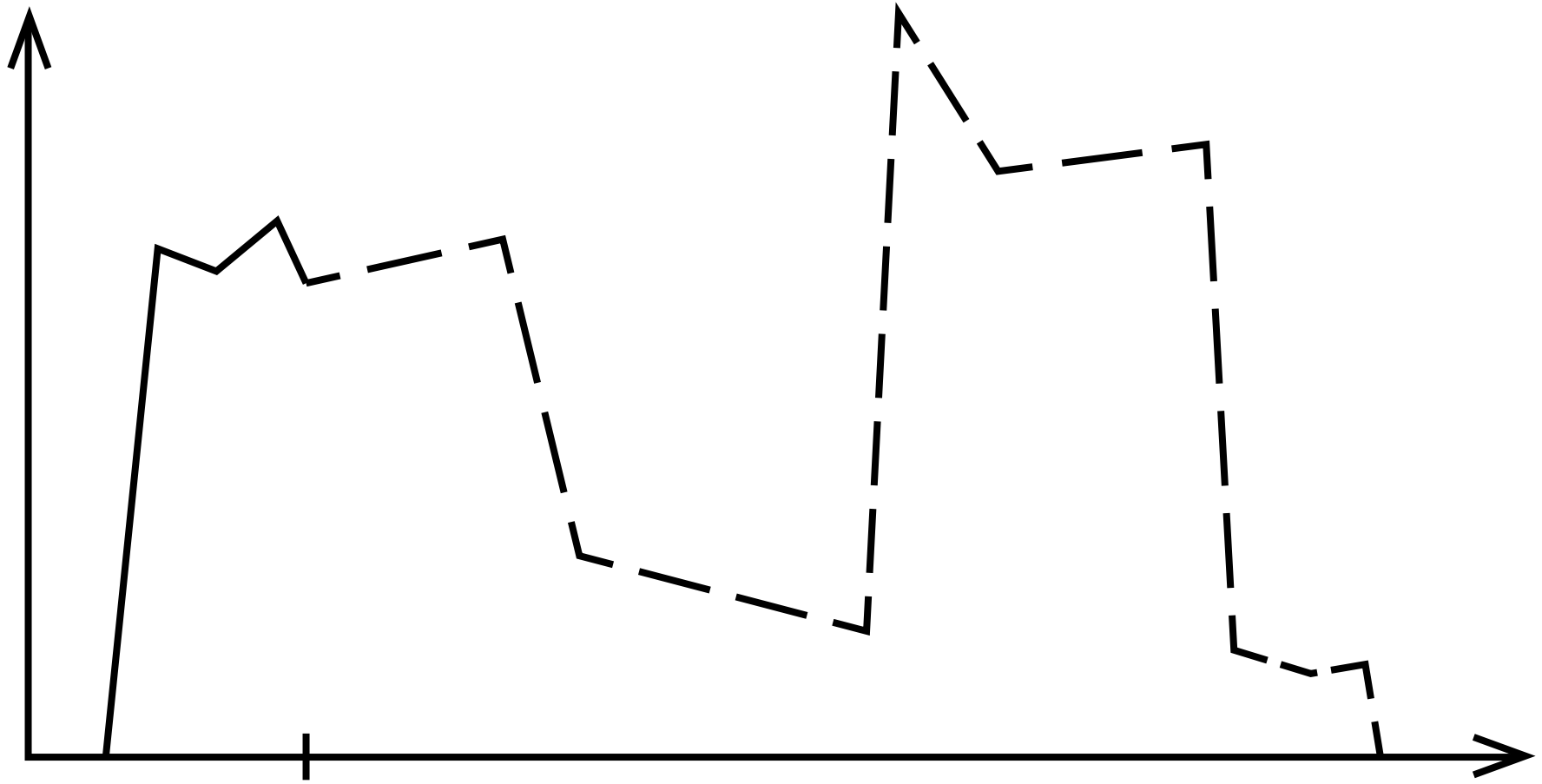
- Can we predict the (inter)arrival times of individual jobs multiple steps ahead?
- To be answered: December 15
- Spoiler: səʌ

# Question

- Can we predict the transient resource usage of individual tasks all the way until their completion?

# Example

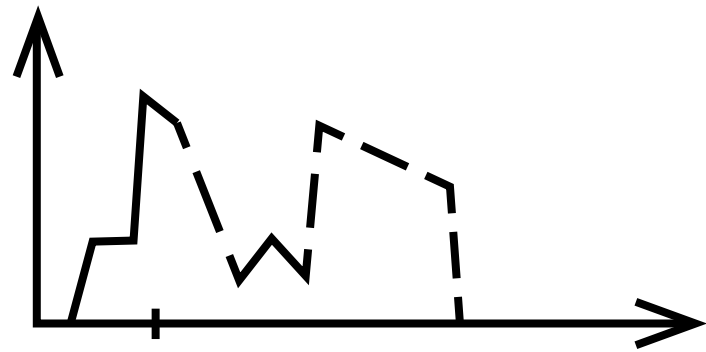
Resource



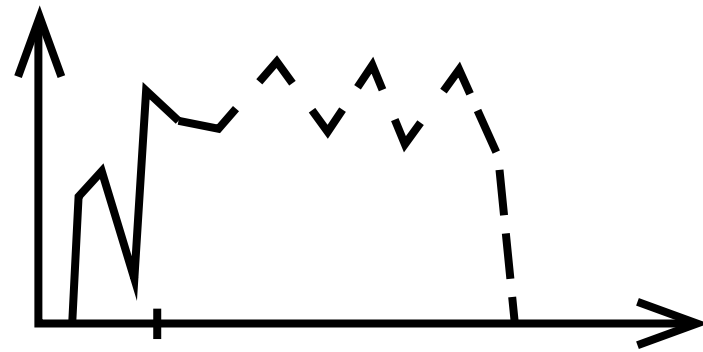
Time

Now

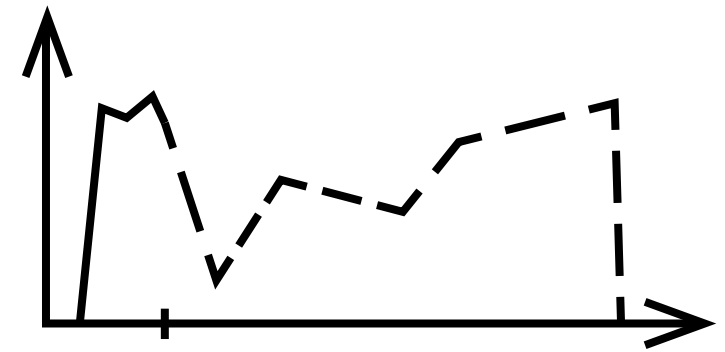
# Example



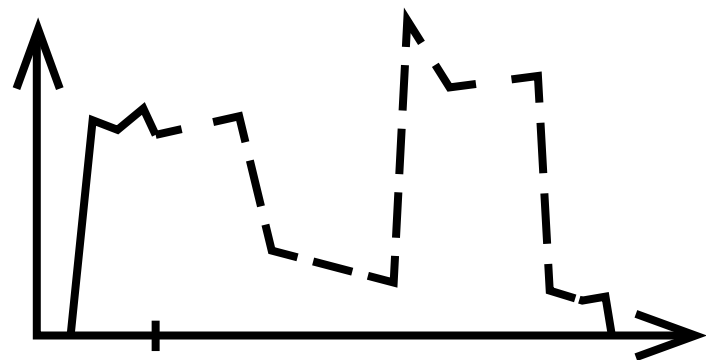
Machine 1



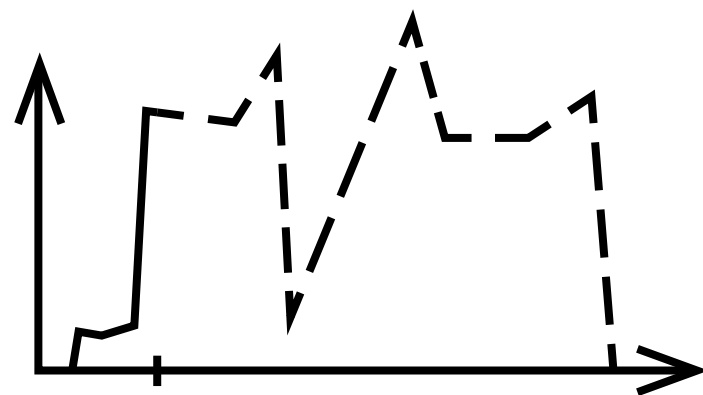
Machine 2



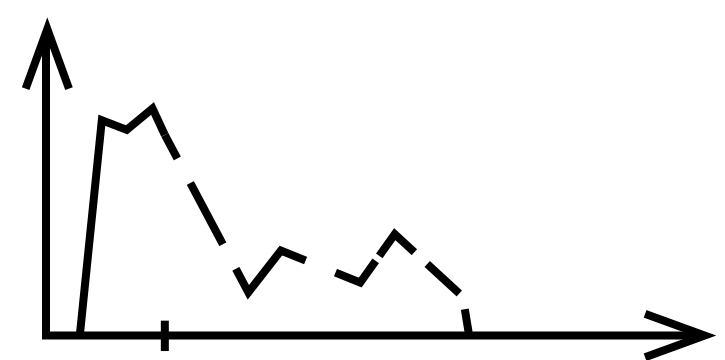
Machine 3



Machine 4



Machine 5



Machine 6

# Google's Resource Usage

- CPU, memory, and disk I/O
- Mean and max over 5-minute intervals
- 300 measurements per interval

# Google's Resource Usage

- 200 GB
- 1,300,000,000 records
- 10 minutes for "SELECT COUNT(\*) FROM table"

# Google's Resource Usage

- Slice and dice
- Organize in a catalog of traces
- Mean CPU (over 5-minute intervals)

Prediction



# Google's Resource Usage

- Sequential data
- Non-trivial structure
- Long-range dependence

# Machine Learning

- No hand-crafted algorithms
- Embrace and learn from experience

# Neural Networks

- The state-of-the-art in you-name-it

# ESWEEK'16

- **Neural Network** Transformation and Co-design under Neuromorphic Hardware Constraints

# ESWEEK'16

- Cambricon: An Instruction Set Architecture for **Neural Networks**

# ESWEEK'16

- RRAM Based **[Deep]** Learning Acceleration

# ESWEEK'16

- Hybrid Network-on-Chip Architectures for Accelerating **Deep Learning** Kernels on Heterogeneous Manycore Platforms

# ESWEEK'16

- CaffePresso: An Optimized Library for **Deep Learning** on Embedded Accelerator-based platforms



# ESWEEK'16

- Going Deeper than **Deep Learning** for Massive Data Analytics under Physical Constraints

# ESWEEK'16

- Zero and Data Reuse-aware Fast Convolution for **Deep Neural Networks** on GPU

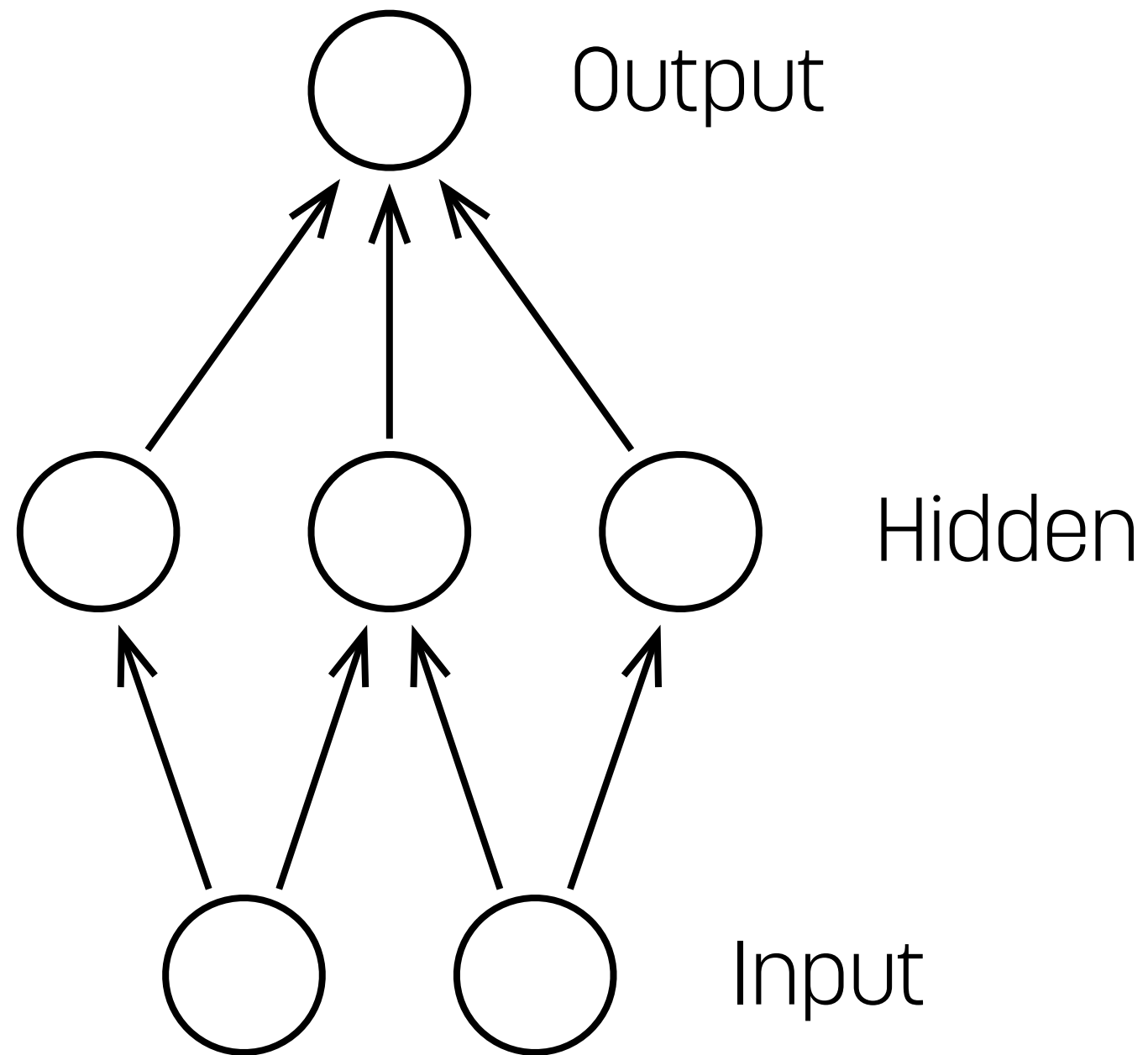
# ESWEEK'16

- Runtime Configurable **Deep Neural Networks** for Energy-Accuracy Trade-off

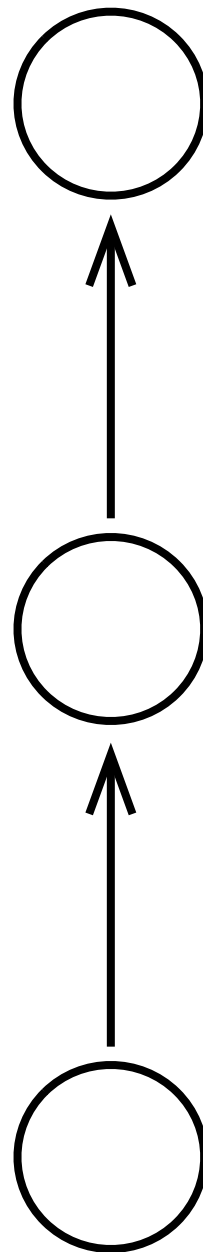
# Neural Networks

- Hot

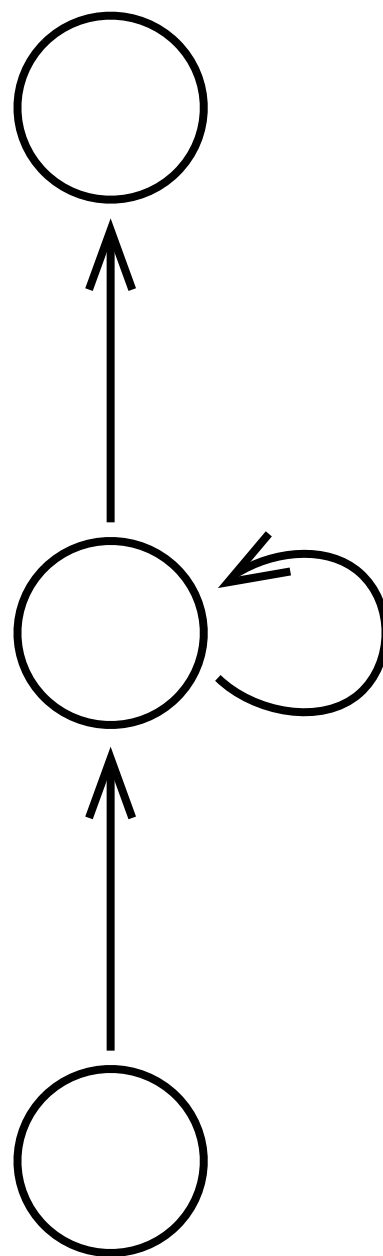
# Feedforward Neural Networks



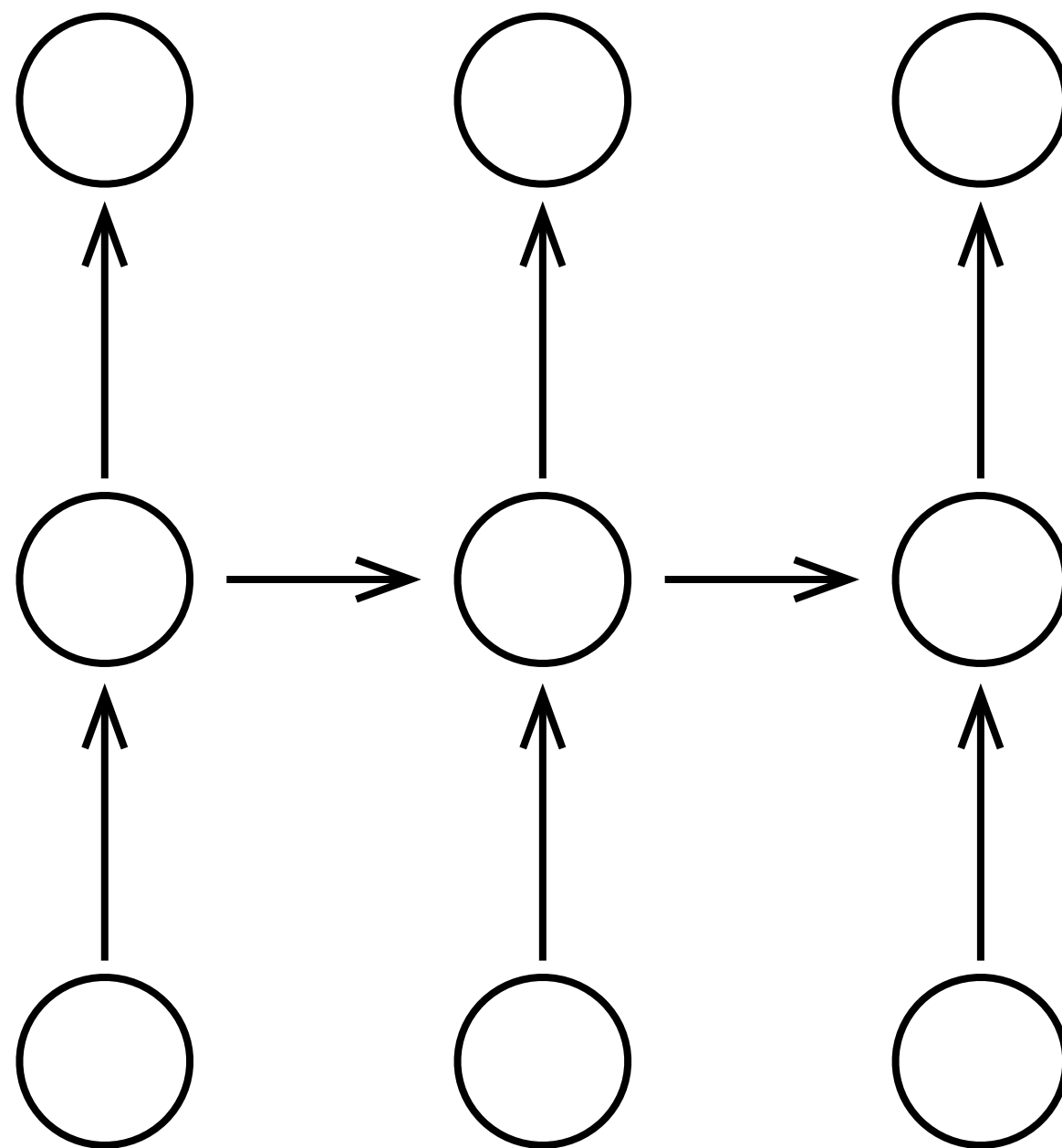
# Feedforward Neural Networks



# Recurrent Neural Networks

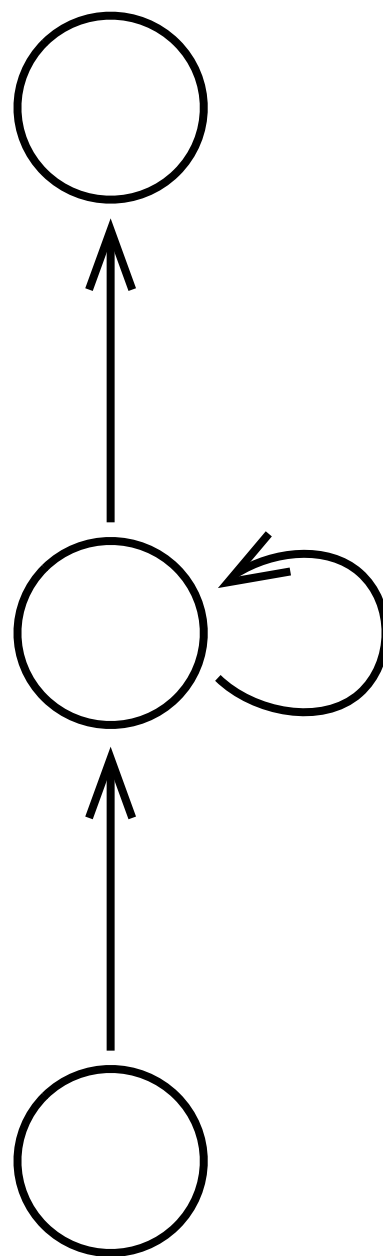


# Recurrent Neural Networks

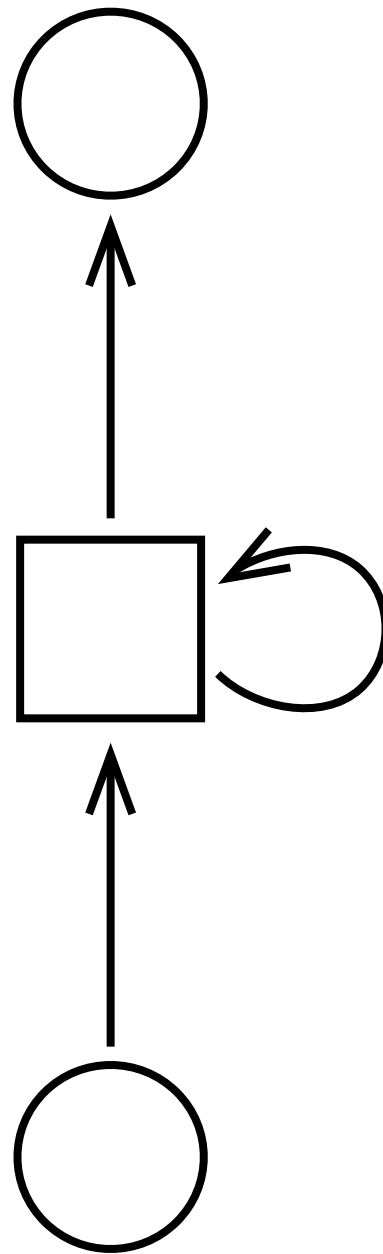




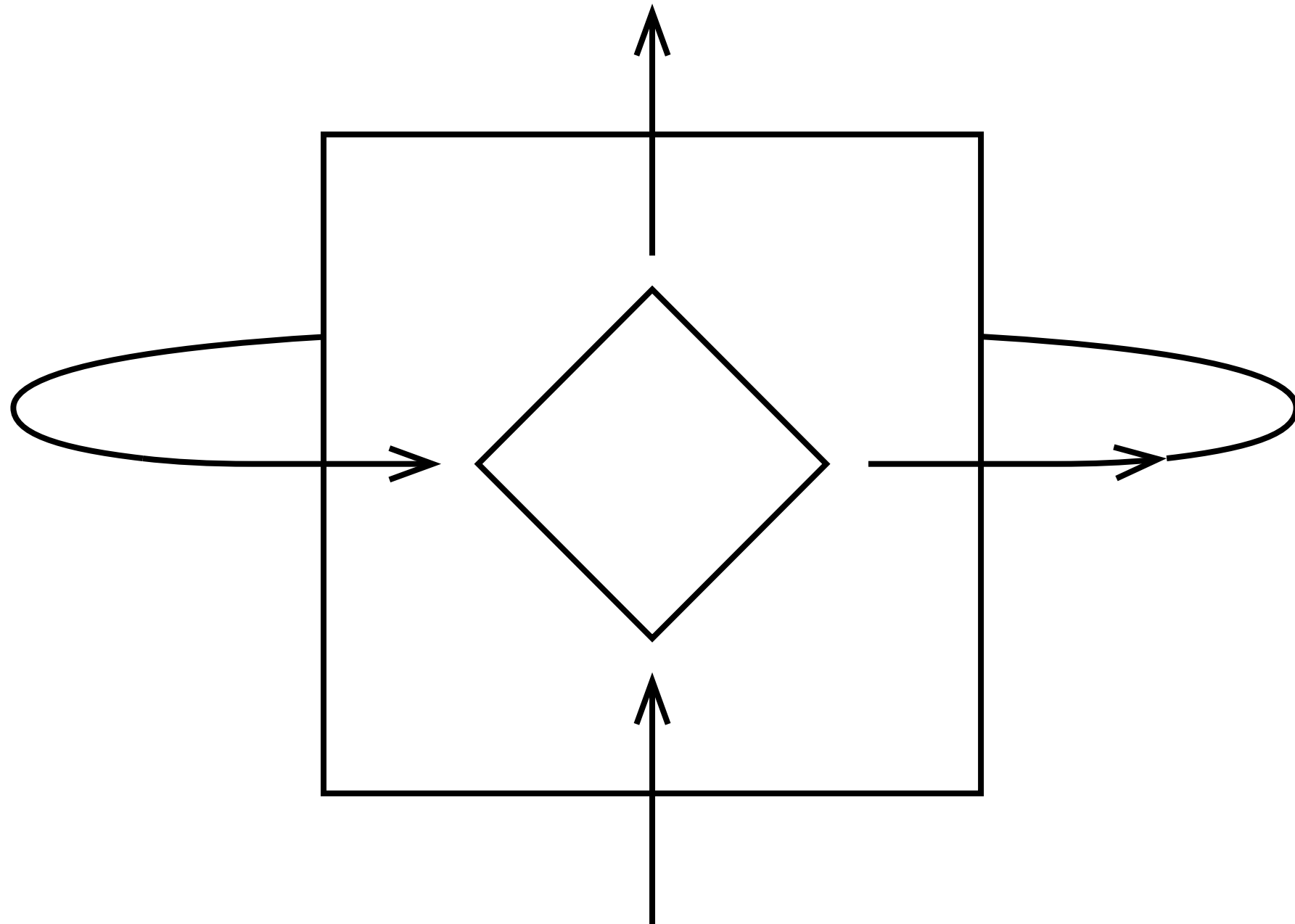
# Recurrent Neural Networks



# Memory Modeling



# Memory Cell



Stay tuned for more...

Thank you!  
Questions?